QÜESTIIÓ, vol. 26, 1-2, p. 87-108, 2002

ESTIMACIÓN DE LA FUNCIÓN DE DISTRIBUCIÓN SOBRE POBLACIONES FINITAS MEDIANTE DISEÑOS MUESTRALES BIETÁPICOS APROPIADOS

J. A. MAYOR GALLEGO
M. MARTÍNEZ BLANES
Universidad de Sevilla*

Con el objeto de estimar la función de distribución de una variable de estudio sobre una población finita, se propone en este trabajo emplear el estimador de Horvitz-Thompson, lo que proporciona una estrategia muestral insesgada, siendo la varianza de dicho estimador una función real de variable real cuya minimización permite obtener diseños muestrales óptimos bajo diferentes criterios.
En este trabajo empleamos la norma $\|\cdot\|_1$ como criterio de optimización, minimizando la norma de la varianza, como función de la matriz del diseño muestral. De esta forma, suponiendo muestreo por conglomerados en dos etapas y considerando como dominio de búsqueda el conjunto de los diseños muestrales de tipo uniforme, en el sentido de ser iguales las probabilidades de inclusión de primer orden, se estudia la obtención de diseños muestrales adecuados para dicha estimación.

Estimating distributions functions using appropriate sampling designs in two stage cluster sampling

Palabras clave: Muestreo, poblaciones finitas, diseño muestral, función de distribución, conglomerados

Clasificación AMS (MSC 2000): 62D05

* Dpto. de Estadística e Investigación Operativa. Universidad de Sevilla. Facultad de Matemáticas. c/ Tarfia s/n, 41012 Sevilla, España. E-mail:mayor@cica.es

– Recibido en marzo de 2001.
– Aceptado en diciembre de 2001.

1. INTRODUCCIÓN

Si bien la teoría del muestreo en poblaciones finitas se ha centrado clásicamente en la estimación de parámetros poblacionales de tipo puntual como totales, medias, proporciones y varianzas, existen una serie de parámetros de tipo funcional que pueden proporcionarnos información relevante acerca del comportamiento global de la población.

En este trabajo consideramos el problema de la estimación de un parámetro de este tipo, en concreto, la función de distribución poblacional asociada a una variable numérica definida sobre la población. Este problema es importante por el interés intrínseco del parámetro funcional mencionado y también por su relación con otros parámetros de tipo no funcional como la mediana, los cuantiles o el índice de Gini, habiendo sido tratado con diferentes enfoques por varios autores.

Así, en relación a la estimación de la mediana y los cuantiles, es obligado citar el trabajo inicial de Wooddruff (1952), en el que se construye un intervalo de confianza para la estimación de la mediana poblacional y otras medidas de posición, empleando el muestreo aleatorio simple. Por otra parte, Sendransk y Meyer (1978) estudian este problema bajo un enfoque puramente probabilístico de distribución de estadísticos ordenados, para muestreo aleatorio simple y estratificado. Hill (1968) emplea un enfoque bayesiano y Kuk y Mak (1989), técnicas de información auxiliar proporcionada por otras variables.

Para el problema de la estimación de la función de distribución poblacional propiamente dicha, también encontramos diferentes enfoques en la bibliografía. Por ejemplo Chambers y Dunstan (1986) emplean un modelo de superpoblación para desarrollar un procedimiento de estimación. Kuk (1988) estudia y compara varios estimadores de la función de distribución poblacional empleando muestreo con probabilidades variables y Rao, Kovar y Mantel (1990) desarrollan estimadores que emplean información auxiliar. Citemos también los trabajos de Chambers *et al.* (1992) y Rao (1994).

A continuación vamos a considerar este problema con un enfoque distinto de los anteriores, y que se basa en el estudio de la función varianza del estimador de Horvitz-Thompson con el fin de buscar diseños muestrales óptimos, en una clase especial de diseños muestrales. Para ello, vamos a considerar una población finita, $U = \{1, 2 \dots N\}$, y sea Y una variable de estudio numérica, cuyos valores sobre U, son $(Y_1, Y_2 \dots Y_N)$, que, sin pérdida de generalidad, supondremos ordenados de menor a mayor, esto es, $Y_1 \leq Y_2 \leq \cdots \leq Y_N$. Nuestro objetivo es la estimación de la función de distribución de la variable Y,

$$F(t) = \frac{1}{N} \mathrm{CARD}(\{i \in U \,|\, Y_i \leq t\})$$

Si empleamos las funciones indicadoras de los intervalos de la forma $[Y_i, +\infty)$, denotándolas $I_{[Y_i,+\infty)}(t)$, podemos expresar $F(t)$ como,

$$F(t) = \frac{1}{N} \sum_{i \in U} I_{[Y_i,+\infty)}(t)$$

es decir, una expresión lineal, que puede ser estimada mediante el estimador de Horvitz-Thompson.

Sea pues m una muestra obtenida de la población U mediante un diseño muestral, $d = (\mathcal{M}, P(\cdot))$, sin reemplazamiento y con matriz de diseño que denotaremos como $\Pi = \{\pi_{ij}\}_{1 \leq i,j \leq N}$, con $\pi_{ii} = \pi_i$. El mencionado estimador de $F(t)$ resulta ser,

$$\widehat{F}(t) = \frac{1}{N} \sum_{i \in m} \frac{I_{[Y_i,+\infty)}(t)}{\pi_i}$$

Como sabemos, este estimador es insesgado y su varianza puede ser expresada mediante la clásica fórmula,

$$V[\widehat{F}(t)] = \frac{1}{N^2} \sum_{i,j \in U} (\pi_{ij} - \pi_i \pi_j) \frac{I_{[Y_i,+\infty)}(t)}{\pi_i} \frac{I_{[Y_j,+\infty)}(t)}{\pi_j}$$

y así, para cada valor de $t \in \mathbb{R}$, podemos emplear dicha expresión como medida de la bondad de la estimación.

2. MUESTREO POR CONGLOMERADOS EN DOS ETAPAS

Al ser la varianza una función, no tiene sentido hablar de varianza mínima en el sentido usual, por ello, vamos a definir un criterio apropiado que nos permita obtener propiedades de los diseños muestrales más adecuados para la estimación que estamos realizando. El criterio que definimos viene determinado por la siguiente distancia, basada en la norma funcional $\|\cdot\|_1$,

$$\mathrm{d}(V[\widehat{F}(t)], 0) = \|V[\widehat{F}(t)]\|_1 = \int_{Y_1}^{Y_N} |V[\widehat{F}(t)]| dt = \int_{Y_1}^{Y_N} V[\widehat{F}(t)] dt$$

Notemos que $V[\widehat{F}(t)]$ cumple los requerimientos analíticos para que esta distancia esté bien definida. En particular, se verifica que $\|V[\widehat{F}(t)]\|_1 = 0$ si y solamente si $V[\widehat{F}(t)] =$

$0, \forall t \in [Y_1, Y_N]$. Notemos también que es posible emplear otro tipo de normas. La elección de la norma $\|\cdot\|_1$ en este trabajo viene justificada por razones de simplicidad en los desarrollos.

Supondremos que se realiza un muestreo por conglomerados en una etapa, así la población sobre la que se realiza el muestreo es una población de conglomerados, $U_c = \{C_1, \ldots, C_i, \ldots, C_M\}$, cada uno con tamaños respectivos $N_1, \ldots, N_i, \ldots, N_M$. Sobre esta población se emplea un diseño muestral $d_c = (\mathcal{M}_c, P_c)$, con matriz de diseño $\{\pi^c_{ij}\}$, siendo m_c la muestra de conglomerados obtenida. En una segunda etapa supondremos que en el conglomerado $C_i,\ i \in m_c$ se emplea un diseño muestral $d_i = (\mathcal{M}_i, P_i)$, con matriz de diseño $\{\pi^i_{kl}\}$, para obtener una muestra m_i con n_i unidades finales.

Las probabilidades de inclusión de dichas unidades finales, se construirán pues a partir de las probabilidades de inclusión de los diseños muestrales involucrados. Dichas probabilidades vienen dadas por,

$$\begin{aligned}
&\pi_k = \pi^c_i \pi^i_k && \text{si} \quad k \in C_i && \forall\, k \in U \\
&\pi_{kl} = \pi^c_i \pi^i_{kl} && \text{si} \quad k, l \in C_i,\ k \neq l && \forall\, k, l \in U \\
&\pi_{kl} = \pi^c_{ij} \pi^i_k \pi^j_l && \text{si} \quad k \in C_i,\ l \in C_j,\ i \neq j && \forall\, k, l \in U
\end{aligned}$$

De esta forma, denotando por m la muestra de unidades finales provenientes de dichos conglomerado, el estimador de Horvitz-Thompson será,

$$\begin{aligned}
\widehat{F}(t) &= \sum_{k \in m} \frac{1}{N} \frac{I_{[Y_k, +\infty)}(t)}{\pi_k} = \sum_{i \in m_c} \sum_{k \in m_i} \frac{1}{N} \frac{I_{[Y_k, +\infty)}(t)}{\pi^c_i \pi^i_k} \\
&= \sum_{i \in m_c} \frac{1}{\pi^c_i} \sum_{k \in m_i} \frac{1}{N} \frac{I_{[Y_k, +\infty)}(t)}{\pi^i_k} = \sum_{i \in m_c} \frac{\widehat{F}_i(t)}{\pi^c_i}
\end{aligned}$$

donde $F_i(t)$ denota el parámetro funcional que se está estimando, valorado sobre el conglomerado C_i, es decir,

$$F_i(t) = \sum_{k \in C_i} \frac{1}{N} I_{[Y_k, +\infty)}(t)$$

y $\widehat{F}_i(t)$ denota su estimación de Horvitz-Thompson, esto es,

$$\widehat{F}_i(t) = \sum_{k \in m_i} \frac{1}{N} \frac{I_{[Y_k, +\infty)}(t)}{\pi^i_k}$$

Observemos que las funciones $F_i(t)$ aparecen como resultado de la descomposición usual del estimador $\widehat{F}(t)$, y no representan las funciones de distribución de los conglomerados.

La varianza de dicho estimador vendrá dada por la expresión usual para un muestreo bietápico, véase Fernández y Mayor (1995), y que adaptada a la estructura poblacional considerada, resulta ser,

$$\begin{aligned} V[\widehat{F}(t)] &= \sum_{i,j\in U_c}\sum (\pi^c_{ij} - \pi^c_i\pi^c_j)\frac{F_i(t)}{\pi^c_i}\frac{F_j(t)}{\pi^c_j} + \sum_{i\in U_c}\frac{1}{\pi^c_i}V[\widehat{F}_i(t)] \\ &= \sum_{i,j\in U_c}\sum (\pi^c_{ij} - \pi^c_i\pi^c_j)\frac{F_i(t)}{\pi^c_i}\frac{F_j(t)}{\pi^c_j} \\ &+ \frac{1}{N^2}\sum_{i\in U_c}\frac{1}{\pi^c_i}\sum_{k,l\in C_i}\sum (\pi^i_{kl} - \pi^i_k\pi^i_l)\frac{I_{[Y_k,+\infty)}(t)}{\pi^i_k}\frac{I_{[Y_l,+\infty)}(t)}{\pi^i_l} = V_1(t) + V_2(t) \end{aligned}$$

Observemos que la varianza se descompone en dos sumandos, $V_1(t)$ y $V_2(t)$, cada uno de los cuales representa el error de muestreo inherente a ambas etapas. Para aplicar nuestra metodología, vamos a estudiar la norma $\|\cdot\|_1$ de esta varianza, bajo ciertas hipótesis realizadas sobre los diseños muestrales. En primer lugar, notemos que al ser $V_1(t) \geq 0$ y $V_2(t) \geq 0$, $\forall t \in \mathbb{R}$, se verifica,

$$\|V[\widehat{F}(t)]\|_1 = \|V_1(t)\|_1 + \|V_2(t)\|_1 \quad \forall t \in \mathbb{R}$$

y en particular $\forall t \in [Y_1, Y_N]$. Seguidamente exponemos una serie de definiciones y resultados relacionados con el desarrollo de las cantidades anteriores.

Definición 1. *Dado un diseño muestral, $d = (\mathcal{M}, P)$, diremos que es uniforme si las probabilidades de inclusión de primer orden son iguales, es decir, $\pi_i = \alpha$, $\forall i$.*

Lema 1. *Si un diseño muestral, $d = (\mathcal{M}, P)$, definido sobre una población U con N unidades, es uniforme y sus muestras son de tamaño fijo, n, entonces se verifica $\pi_i = n/N, \forall i \in U$.*

Demostración

Dado $i \in U$, sea H_i una variable aleatoria con distribución de Bernoulli, definida sobre $\mathcal{M}$ como $H_i(m) = 1$ si $i \in m$ y $H_i(m) = 0$ si $i \neq m$, $\forall m \in \mathcal{M}$. La esperanza matemática de H_i es $E[H_i] = \pi_i$. Se tiene entonces,

$$\sum_{i\in U} H_i = n \quad \Rightarrow \quad \sum_{i\in U} \pi_i = \sum_{i\in U} E[H_i] = n$$

de donde se deduce el resultado. □

Observemos que el diseño muestral aleatorio simple, es decir, el formado por todas las posibles muestras de tamaño fijo n, con probabilidades iguales, es decir, $P(m) = 1/\binom{N}{n}$, es un diseño uniforme de tamaño fijo. Denotaremos este diseño muestral como $\mathrm{MAS}(N,n)$.

Lema 2. *Si los diseños muestrales $d_i,\ i \in U_c$ son uniformes con tamaños fijos respectivos n_i, y si denotamos,*

$$V^{(i)}(t) = \sum_{k,l\in C_i}\sum (\pi^i_{kl} - \pi^i_k\pi^i_l)\frac{I_{[Y_k,+\infty)}(t)}{\pi^i_k}\frac{I_{[Y_l,+\infty)}(t)}{\pi^i_l} \qquad i \in U_c$$

se verifica,

$$||V^{(i)}(t)||_1 = \frac{1}{2}\sum_{k,l\in C_i}\sum |Y_k - Y_l| - \frac{N_i^2}{2n_i^2}\sum_{k,l\in C_i}\sum \pi^i_{kl}|Y_k - Y_l|$$

Si además, d_i =MAS(N_i,n_i), $\forall i \in U_c$, entonces,

$$||V^{(i)}(t)||_1 = \frac{N_i - n_i}{2(N_i-1)n_i}\sum_{k,l\in C_i}\sum |Y_k - Y_l|$$

Demostración

Observemos que dados $k,l \in C_i$, se tiene,

$$\int_{Y_1}^{Y_N} I_{[Y_k,+\infty)}(t) I_{[Y_l,+\infty)}(t)\,dt = \int_{Y_1}^{Y_N} I_{[\max\{Y_k,Y_l\},+\infty)}(t)\,dt = Y_N - \max\{Y_k,Y_l\}$$

por consiguiente,

$$||V^{(i)}(t)||_1 \quad = \int_{Y_1}^{Y_N} V^{(i)}(t)\,dt = \sum_{k,l\in C_i}\sum \frac{\pi^i_{kl} - \pi^i_k\pi^i_l}{\pi^i_k\pi^i_l}\int_{Y_1}^{Y_N} I_{[Y_k,+\infty)}(t) I_{[Y_l,+\infty)}(t)\,dt$$

$$= \sum_{k,l\in C_i}\sum \frac{\pi^i_{kl} - \pi^i_k\pi^i_l}{\pi^i_k\pi^i_l}(Y_N - \max\{Y_k, Y_l\})$$

$$= \sum_{k,l\in C_i}\sum \frac{\pi^i_{kl}}{\pi^i_k\pi^i_l}(Y_N - \max\{Y_k, Y_l\}) - \sum_{k,l\in C_i}\sum (Y_N - \max\{Y_k, Y_l\})$$

Por ser d_i uniforme de tamaño fijo, tendremos por el Lema 1, que $\pi^c_l = n_i/N_i$, $\forall l \in C_i$, siendo pues,

$$\begin{aligned}
&\sum_{k,l\in C_i}\sum \frac{\pi^i_{kl}}{\pi^i_k\pi^i_l}(Y_N - \max\{Y_k, Y_l\}) \\
&\quad = \frac{N_i^2}{n_i^2}\sum_{k,l\in C_i}\sum \pi^i_{kl}(Y_N - \frac{1}{2}(Y_k + Y_l + |Y_k - Y_l|) \\
&\quad = \frac{N_i^2}{n_i^2}\left[\sum_{k,l\in C_i}\sum \pi^i_{kl}Y_N - \frac{1}{2}\sum_{k\in C_i} Y_k \sum_{l\in C_i}\pi^i_{kl} - \frac{1}{2}\sum_{l\in C_i} Y_l \sum_{k\in C_i}\pi^i_{kl} - \frac{1}{2}\sum_{k,l\in C_i}\sum \pi^i_{kl}|Y_k - Y_l|\right] \\
&\quad = \frac{N_i^2}{n_i^2}\left[n_i^2 Y_N - \frac{n_i^2}{N_i}T_i(Y) - \frac{1}{2}\sum_{k,l\in C_i}\sum \pi^i_{kl}|Y_k - Y_l|\right] \\
&\quad = N_i^2 Y_N - N_i T_i(Y) - \frac{N_i^2}{2n_i^2}\sum_{k,l\in C_i}\sum \pi^i_{kl}|Y_k - Y_l|
\end{aligned}$$

donde hemos denotado $T_i(Y) = \sum_{k\in C_i} Y_k$, y hemos tenido en cuenta que por ser el diseño muestral uniforme y de tamaño de muestra fijo, verifica,

$$\sum_{k\in C_i}\pi^i_{kl} = n_i\pi^i_l = \frac{n_i^2}{N_i} \qquad \text{y} \qquad \sum_{k,l\in C_i}\sum \pi^i_{kl} = n_i^2$$

Por otra parte, un cálculo similar al anterior proporciona,

$$\sum_{k,l\in C_i}\sum (Y_N - \max\{Y_k, Y_l\}) = N_i^2 Y_N - N_i T_i(Y) - \frac{1}{2}\sum_{k,l\in C_i}\sum |Y_k - Y_l|$$

Restando ambas se obtiene inmediatamente el primer resultado. En caso de ser $d_i = \text{MAS}(N_i, n_i)$, $i \in U_c$, se tendrá además para las probabilidades de inclusión de segundo

orden $\pi^i_{kl} = n_i(n_i-1)/N_i(N_i-1),\ k \neq l$, y basta sustituir para obtener, mediante un cálculo directo, el segundo resultado. □

Lema 3. *Las funciones $F_i(t)$ definidas anteriormente, verifican,*

$$\int_{Y_1}^{Y_N} F_i(t)\,F_j(t)\,dt = \frac{1}{N^2}\sum_{k\in C_i,l\in C_j}\sum\sum (Y_N - \max\{Y_k,Y_l\}) \qquad \forall i,j \in U_c$$

Demostración

En primer lugar, observemos que,

$$F_i(t)\,F_j(t) = \frac{1}{N^2}\sum_{k\in C_i,l\in C_j}\sum\sum I_{[Y_k,+\infty)}(t) I_{[Y_l,+\infty)}(t)$$

luego,

$$\int_{Y_1}^{Y_N} F_i(t)\,F_j(t)\,dt = \frac{1}{N^2}\sum_{k\in C_i,l\in C_j}\sum\sum \int_{Y_1}^{Y_N} I_{[Y_k,+\infty)}(t) I_{[Y_l,+\infty)}(t)\,dt$$

y basta sustituir la expresión obtenida en la demostración del Lema anterior para esta última integral. □

Aplicando los resultados anteriores, obtenemos el siguiente resultado acerca de la varianza de la estimación.

Teorema 4. *Si los diseños muestrales $d_i,\ i \in U_c$, son uniformes con tamaños fijos respectivos n_i, se verifica,*

$$\begin{aligned}
\|V[\widehat{F}(t)]\|_1 \;=\; & \frac{1}{N^2}\sum_{i,j\in U_c}\sum\frac{\pi^c_{ij}}{\pi^c_i\pi^c_j}\sum_{k\in C_i,l\in C_j}\sum\sum (Y_N - \max\{Y_k,Y_l\}) \\
& -\frac{1}{N^2}\sum_{i,j\in U_c}\sum\sum_{k\in C_i,l\in C_j}\sum\sum (Y_N - \max\{Y_k,Y_l\}) \\
& +\frac{1}{N^2}\sum_{i\in U_c}\frac{1}{2\pi^c_i}\left(\sum_{k,l\in C_i}\sum\sum |Y_k - Y_l| - \frac{N_i^2}{n_i^2}\sum_{k,l\in C_i}\sum\sum \pi^i_{kl}|Y_k - Y_l|\right)
\end{aligned}$$

Si además d_i =MAS(N_i, n_i), $i \in U_c$, el último sumando se puede expresar como,

$$\frac{1}{N^2} \sum_{i \in U_c} \frac{N_i}{\pi_i^c} \left(\frac{N_i}{n_i} - 1 \right) \frac{1}{2N_i(N_i - 1)} \sum_{k,l \in C_i} \sum |Y_k - Y_l|$$

□

3. REDUCCIÓN DE LA VARIANZA

Aunque la expresión obtenida permite considerar la búsqueda de diseños uniformes en cada uno de los conglomerados, que reduzcan la varianza, aquí restringiremos el planteamiento suponiendo d_i =MAS(N_i, n_i), $i \in U_c$, es decir, muestreo aleatorio en los conglomerados.

Como puede verse en el Teorema 4, bajo esta hipótesis, $\|V[\widehat{F}(t)]\|_1$ se descompone en tres términos, de los cuales el primero y el tercero dependen del diseño muestral empleado para obtener la muestra de conglomerados. Por otra parte, la cantidad que aparece en dicha expresión,

$$\frac{1}{2N_i(N_i - 1)} \sum_{k,l \in C_i} \sum |Y_k - Y_l|$$

puede considerarse como una medida de la dispersión de la variable de estudio en el conglomerado $i-$ésimo.

De esta forma, si suponemos afijación proporcional en los conglomerados, de manera que el tamaño de muestra en cada uno de ellos se realiza de forma proporcional al tamaño del mismo, y suponemos además que los conglomerado presentan una dispersión similar con respecto a la variable de estudio, las cantidades,

$$\left(\frac{N_i}{n_i} - 1 \right) \frac{1}{2N_i(N_i - 1)} \sum_{k,l \in C_i} \sum |Y_k - Y_l|$$

son muy homogéneas, luego una forma de disminuir la aportación del último término es tomar probabilidades de inclusión de primer orden proporcionales a los tamaños de los conglomerados, es decir, $\pi_i^c = nN_i/N$. Se obtiene entonces el siguiente resultado.

Teorema 5. *Si el muestreo en los conglomerados se realiza mediante diseños muestrales aleatorios simples, y si las probabilidades de inclusión de primer orden asociadas al muestreo de conglomerados vienen dadas por $\pi_i^c = nN_i/N$, $i \in U_c$, el primer sumando de la expresión de $\|V[\widehat{F}]\|_1$, obtenida en el Teorema 4. puede ser expresado como,*

$$\frac{1}{nN}\sum_{i\in U_c} N_i\left(Y_N-\overline{Y}_i-\frac{1}{2N_i^2}\sum_{k,l\in C_i}\sum|Y_k-Y_l|\right)$$

$$+\frac{1}{n^2}\left[n(n-1)(Y_N-\overline{Y})-\frac{1}{2}\sum_{i,j\in U_c,i\neq j}\sum\frac{\pi_{ij}^c}{N_iN_j}\sum_{k\in C_i,l\in C_j}\sum|Y_k-Y_l|\right]$$

donde $\overline{Y}$ denota la media poblacional de la variable de estudio, é $\overline{Y}_i$ la media sobre el conglomerado $i-$ésimo.

Demostración

En primer lugar observemos que dados $i,j\in U_c$, se tiene,

$$\begin{aligned}\sum_{k\in C_i,l\in C_j}\sum(Y_N-\max\{Y_k,Y_l\}) &= \sum_{k\in C_i,l\in C_j}\sum\left(Y_N-\frac{1}{2}Y_k-\frac{1}{2}Y_l-\frac{1}{2}|Y_k-Y_l|\right)\\ &= N_iN_jY_N-\frac{1}{2}N_jT_i(Y)-\frac{1}{2}N_iT_j(Y)-\frac{1}{2}\sum_{k\in C_i,l\in C_j}\sum|Y_k-Y_l|\\ &= N_iN_j\left(Y_N-\frac{1}{2}(\overline{Y}_i+\overline{Y}_j)-\frac{1}{2N_iN_j}\sum_{k\in C_i,l\in C_j}\sum|Y_k-Y_l|\right)\end{aligned}$$

así pues,

$$\begin{aligned}&\sum_{i,j\in U_c}\sum\frac{\pi_{ij}^c}{\pi_i^c\pi_j^c}\sum_{k\in C_i,l\in C_j}\sum(Y_N-\max\{Y_k,Y_l\})\\ &\quad=\frac{N}{n}\sum_{i\in U_c}N_i\left(Y_N-\overline{Y}_i-\frac{1}{2N_i^2}\sum_{k,l\in C_i}\sum|Y_k-Y_l|\right)\\ &\qquad+\frac{N^2}{n^2}\sum_{i,j\in U_c,i\neq j}\sum\pi_{ij}^c\left(Y_N-\frac{1}{2}(\overline{Y}_i+\overline{Y}_j)-\frac{1}{2N_iN_j}\sum_{k\in C_i,l\in C_j}\sum|Y_k-Y_l|\right)\\ &\quad=\frac{N}{n}\sum_{i\in U_c}N_i\left(Y_N-\overline{Y}_i-\frac{1}{2N_i^2}\sum_{k,l\in C_i}\sum|Y_k-Y_l|\right)\\ &\qquad+\frac{N^2}{n^2}\left[n(n-1)Y_N-n(n-1)\overline{Y}-\frac{1}{2}\sum_{i,j\in U_c,i\neq j}\sum\frac{\pi_{ij}^c}{N_iN_j}\sum_{k\in C_i,l\in C_j}\sum|Y_k-Y_l|\right]\end{aligned}$$

de donde se deduce inmediatamente el resultado. Nótese que hemos empleado la relación,

$$\sum_{j\in U_c, j\neq i} \pi^c_{ij} = (n-1)\pi^c_i = \frac{n(n-1)N_i}{N}$$

□

El resultado anterior nos dice que podemos reducir la varianza de la estimación actuando sobre el último término, es decir, eligiendo adecuadamente las probabilidades de inclusión de segundo orden, bajo el supuesto de que las de primer orden son proporcionales a los tamaños de los conglomerados. No obstante, bajo la hipótesis de que dichos tamaños son similares, podemos considerar diseño uniforme para el muestreo de los mismos. En este sentido se tiene el siguiente resultado, cuya demostración, similar a la del anterior, omitimos.

Teorema 6. *Si el muestreo en los conglomerados se realiza mediante diseños muestrales aleatorios simples, y si la selección de conglomerados se realiza mediante un diseño muestral uniforme, entonces el primer sumando de la expresión de* $\|V[\widehat{F}]\|_1$*, obtenida en el Teorema 4 puede expresarse como,*

$$\begin{aligned}&\frac{M}{n}\sum_{i\in U_c} N_i^2\left(Y_N - \overline{Y}_i - \frac{1}{2N_i^2}\sum_{k,l\in C_i}\sum |Y_k - Y_l|\right)\\ &\quad+\frac{M^2}{n^2}\sum_{i,j\in U_c, i\neq j}\sum \pi^c_{ij}N_iN_j\left(Y_N - \frac{1}{2}(\overline{Y}_i+\overline{Y}_j) - \frac{1}{2N_iN_j}\sum_{k\in C_i, l\in C_j}\sum |Y_k - Y_l|\right)\end{aligned}$$

donde $\overline{Y}_i$ *denota la media de la variable de estudio sobre el conglomerado* $i-$*ésimo.* □

Según este último resultado, y con las hipótesis consideradas, podemos reducir la varianza de la estimación minimizando la cantidad,

$$\begin{aligned}&\sum_{i,j\in U_c, i\neq j}\sum \pi^c_{ij}N_iN_j\left(Y_N - \frac{1}{2}(\overline{Y}_i+\overline{Y}_j) - \frac{1}{2N_iN_j}\sum_{k\in C_i, l\in C_j}\sum |Y_k - Y_l|\right)\\ &\qquad=\sum_{i,j\in U_c, i\neq j}\sum \pi^c_{ij}\sum_{k\in C_i, l\in C_j}\sum (Y_N - \max\{Y_k, Y_l\})\end{aligned}$$

en las variables π_{ij}^c. Observemos que bajo la hipótesis de homogeneidad en los tamaños de los conglomerados, dicha minimización tenderá a asignar probabilidades de inclusión de segundo orden mayores para pares de conglomerados más distantes en el sentido de ser mayor la cantidad,

$$\frac{1}{2N_iN_j}\sum\sum_{k\in C_i,l\in C_j}|Y_k-Y_l|$$

ya que dicha cantidad se puede considerar como una medida de la disparidad entre los conglomerados C_i y C_j, en relación a los valores de la variable de estudio sobre ellos, así, las parejas de conglomerados más dispares deberían tener probabilidades de inclusión de segundo orden más elevadas.

Notemos finalmente que la variable Y es desconocida, por lo que el problema de minimización obtenido no puede ser considerado directamente, y por esta razón realizaremos algunas hipótesis sobre la estructura de la población. Así, vamos a estudiar la expresión a minimizar una estructura poblacional muy común, y que formalizaremos mediante un modelo de superpoblación adecuado.

4. MODELO DE REGRESIÓN LINEAL POR CONGLOMERADOS

$$Y_k=\alpha+\beta X_i+\varepsilon_k,\quad \beta>0,\quad E_s[\varepsilon_k]=0\qquad \forall k\in C_i,\ \forall i\in U_c$$

Dicho modelo formaliza aquellas situaciones en las cuales existe una relación aproximada, de tipo lineal entre la variable de estudio y una variable auxiliar, completamente conocida, X, con el mismo valor en cada uno de los conglomerados.

Este tipo de relaciones suelen darse en situaciones reales en las cuales la variable de estudio presenta cierta homogeneidad en cada conglomerado pero estos tienen comportamientos distintos, aunque con un patrón de tipo lineal en relación a una variable conocida sobre U_c. Por ejemplo, en un estudio económico en el cual las provincias españolas son los conglomerados y los municipios las unidades finales, la renta percápita provincial en un determinado año estará en concomitancia con variables de estudio de tipo económico definidas sobre los municipios como renta percápita municipal en años diferentes, volumen económico de transacciones comerciales, etc.

Bajo este modelo, sustituimos el problema planteado por el de minimizar la expresión,

$$E_s\left[\sum\sum_{i,j\in U_c,i\neq j}\pi_{ij}^c\sum\sum_{k\in C_i,l\in C_j}(Y_N-\max\{Y_k,Y_l\})\right]$$

Denotando por X_{MAX} el valor máximo de la variable X, y por ν el índice del conglomerado al que pertenece la unidad poblacional N-ésima, y suponiendo que $k \in C_i$ y $l \in C_j$, se tiene,

$$\begin{aligned} E_s[(Y_N - \max\{Y_k, Y_l\})] &= E_s[Y_N] - E_s[\max\{Y_k, Y_l\}] \\ &\leq \alpha + \beta X_\nu - \max\{E_s[Y_k], E_s[Y_k]\} = \alpha + \beta X_\nu - \max\{\alpha + \beta X_i, \alpha + \beta X_j\} \\ &\leq \beta(X_{\mathrm{MAX}} - \max\{X_i, X_j\}) \end{aligned}$$

así pues,

$$\begin{aligned} E_s \left[\sum_{i,j \in U_c, i \neq j} \sum \pi_{ij}^c \sum_{k \in C_i, l \in C_j} \sum (Y_N - \max\{Y_k, Y_l\}) \right] \\ \leq \beta \sum_{i,j \in U_c, i \neq j} \sum \pi_{ij}^c \sum_{k \in C_i, l \in C_j} \sum (X_{\mathrm{MAX}} - \max\{X_i, X_j\}) \\ = \beta \sum_{i,j \in U_c, i \neq j} \sum \pi_{ij}^c N_i N_j (X_{\mathrm{MAX}} - \max\{X_i, X_j\}) \end{aligned}$$

con lo cual, los diseños muestrales apropiados para la estimación que estamos realizando son aquellos que minimizan la expresión,

$$\sum_{i,j \in U_c, i \neq j} \sum \pi_{ij}^c N_i N_j (X_{\mathrm{MAX}} - \max\{X_i, X_j\})$$

4.1. Implementación bajo el modelo de regresión lineal por conglomerados

A continuación vamos a estudiar la viabilidad de la metodología descrita, en el sentido de su aplicabilidad a situaciones reales, y en el contexto del modelo de superpoblación considerado.

Como se ha visto, es factible utilizar probabilidades de inclusión de primer orden, π_i^c, proporcionales a los tamaños de los conglomerados, si estos presentaran grandes diferencias. No obstante, en este trabajo suponemos que ello no ocurre, por lo que empleamos un diseño uniforme. Con esta elección, tendremos necesariamente $\pi_i^c = n/M,\ \forall i \in U_c$, y la expresión a minimizar es,

$$\sum_{i,j\in U_c, i\neq j}\sum \pi_{ij}^c N_i N_j (X_{\mathrm{MAX}} - \max\{X_i, X_j\})$$

es decir, una función objetivo lineal en las variables $\pi_{ij}^c,\ i \neq j$.

Debido a la simetría de las cantidades π_{ij}^c, tendremos en total $M(M-1)/2$ variables, y teniendo en cuenta que el número de conglomerados no suele ser elevado, en comparación con el tamaño de la población, el problema resultante posee un tamaño razonable para abordar su resolución con los programas usuales. Por ejemplo, para $M = 100$ conglomerados, tendríamos un problema de programación lineal con 4950 variables, abordable por las rutinas de uso común para este tipo de problemas.

En cuanto a las restricciones del problema, tendremos en primer lugar $\pi_{ij}^c > 0$, con objeto de que el diseño permita estimar la varianza de la estimación. Por otra parte, de la relación entre probabilidades de inclusión de primer y segundo orden obtendremos las restricciones,

$$\sum_{j\in U_c, j\neq i} \pi_{ij}^c = (n-1)\pi_i^c = \frac{(n-1)n}{M} \qquad \forall i \in U_c$$

y finalmente, con objeto de poder construir estimaciones no negativas de la varianza de la estimación, introducimos las restricciones,

$$\pi_{ij}^c \leq \pi_i^c \pi_j^c = \frac{n^2}{M^2} \qquad \forall i, j \in U_c,\ i \neq j$$

En el apéndice de este trabajo se exponen algunos aspectos computacionales, incluyendo una rutina de implementación en el lenguaje de descripción de problemas AMPL, Fourer *et al.* (1993), así como un ejemplo numérico sobre varias poblaciones.

5. CONCLUSIONES

En primer lugar, es interesante observar que las metodologías clásicas estudiadas en el muestreo en poblaciones finitas, para estimar los parámetros usuales como medias y totales, no son las más apropiadas para estimar parámetros de tipo funcional como la función de distribución poblacional. En efecto, cuando se estiman parámetros puntuales como medias y totales poblacionales, la reducción de varianza se consigue a través de probabilidades de inclusión de primer orden proporcionales a una variable auxiliar relacionada con la de estudio. Por contra, el la estimación de la función de distribución, este planteamiento carece de sentido por la estructura especial de dicho parámetro.

La metodología introducida en este trabajo, basada en minimizar una determinada norma de la función $V[\widehat{F}(t)]$, en este caso la norma $\|\cdot\|_1$, se manifiesta como una alternativa prometedora pues minimizando dicha norma también disminuimos globalmente dicha función de varianza.

Para el caso de que exista relación entre la variable de estudio y una variable auxiliar, formalizada mediante un modelo de superpoblación del tipo de regresión lineal por conglomerados, la reducción de varianza se realiza a partir de la resolución de un problema de programación lineal que proporciona las probabilidades de inclusión de segundo orden. Los resultados numéricos obtenidos en el estudio empírico que se muestra en el apéndice indican que esta metodología puede producir de forma efectiva una reducción del error de la estimación.

6. APÉNDICE

A continuación, exponemos de forma esquemática una rutina en el lenguaje de descripción de problemas AMPL, para el tratamiento del problema considerado. Obsérvese que la restricción $\pi_{ij}^c > 0$, intratable de forma práctica, ha sido sustituida por $\pi_{ij}^c \geq \varepsilon$, siendo $\varepsilon > 0$ un número real muy próximo a cero.

```
param M;
set U:= 1..M;
set PAIRS:={i in U,j in U: i < j};
param X{U} >= 0;
var PI{PAIRS} >= epsilon <=1;
minimize FUNC: sum{(i,j) in PAIRS} (PI[i,j]*Ni*Nj
                                   *(XMAX-max(X[i],X[j])));
subject to RELATION {i in U} :  sum{j in U : j > i} PI[i,j]
                 + sum{j in U : j < i} PI[j,i] = (n-1)*n/M;
subject to BOUND {(i,j) in PAIRS}: PI[i,j] <= n*n/M*M;
```

Como aplicación, vamos a realizar un estudio empírico comparativo, empleando tres poblaciones, U_1, U_2 y U_3, cada una con $M = 20$ conglomerados, con tamaños iguales $N_i = 100$, $\forall i$. Estas poblaciones son artificiales pero indicativas de distintas situaciones que podemos encontrar en problemas reales, y han sido generadas a partir del modelo de superpoblación considerado.

Para la población U_1, hemos empleado los valores $X_i = 20 + i$, $i = 1, \ldots, 20$, siendo los valores poblacionales generados según el modelo,

$$Y_k = 10 + 2X_i + \varepsilon_k \qquad \forall k \in C_i,\ \forall i \in U_c$$

con $\varepsilon_k \sim N(0, 3^2)$, es decir, normales de esperanza $\mu = 0$ y varianza $\sigma^2 = 3^2$.

Para la población U_2, hemos empleado los valores $X_i = i^2,\ i = 1,\ldots,20$, siendo los valores poblacionales generados según el modelo,

$$Y_k = 20 + 2X_i + \varepsilon_k \qquad \forall k \in C_i,\ \forall i \in U_c$$

con $\varepsilon_k \sim N(0, 5^2)$.

Finalmente, para la población U_3, hemos tomado los valores $X_i,\ i = 1,\ldots,20$, dados por,

$$\{1, 5, 50, 51, 53, 55, 57, 100, 120, 150, 155, 160, 165, 170, 175, 180, 185, 190, 500, 1000\}$$

con los valores poblacionales generados según el modelo,

$$Y_k = 20 + 2X_i + \varepsilon_k \qquad \forall k \in C_i,\ \forall i \in U_c$$

siendo $\varepsilon_k \sim N(0, 10^2)$.

Como puede verse, la primera población presenta valores de la variable de estudio con cierta homogeneidad. La segunda es menos homogénea y la tercera tiene una estructura muy dispar, con ciertos valores extremos muy distantes de la masa principal. En todos los caso, se supone muestreo aleatorio simple, $\mathrm{MAS}(100, 10)$, para el muestreo dentro de los conglomerados, es decir, en cada uno de los conglomerados muestreados en la primera etapa se extraen muestras aleatorias simples de 10 unidades finales.

Vamos a comparar la metodología considerada en este trabajo con el muestreo aleatorio simple de conglomerados, para lo cual utilizaremos como medida de eficiencia relativa la siguiente cantidad, expresada como porcentaje,

$$C = 100 \times \frac{||V[\widehat{F}(t)]||_1^{\mathrm{MAS}} - ||V[\widehat{F}(t)]||_1^{\mathrm{MET}}}{||V[\widehat{F}(t)]||_1^{\mathrm{MAS}}}\,\%$$

siendo $||V[\widehat{F}(t)]||_1^{\mathrm{MAS}}$ y $||V[\widehat{F}(t)]||_1^{\mathrm{MET}}$ respectivamente las normas $||\cdot||_1$ de la varianza expuesta en el Teorema 4, para el muestreo aleatorio simple y para el método estudiado.

Observemos que en $||V[\widehat{F}(t)]||_1^{\mathrm{MET}}$ aparecen las probabilidades de inclusión de segundo orden obtenidas por minimización, mientras que en $||V[\widehat{F}(t)]||_1^{\mathrm{MAS}}$ aparecen las correspondientes al muestreo aleatorio simple de conglomerados, cuyos valores son $\pi_{ij}^c = n(n-1)/M(M-1)\ i \neq j$.

Un coeficiente mayor que cero indicará un aumento de eficiencia con respecto al muestreo aleatorio simple, bajo el criterio considerado, y el correspondiente porcentaje indicará la cuantía de dicho aumento. Los resultados comparativos obtenidos se exponen a continuación para cada una de las tres poblaciones y para tamaños muestrales $n = 3, 4, 5$, siendo n el número de conglomerados muestreados en la primera etapa. Las cantidades reflejadas en esta tabla han sido obtenidas por computación directa a partir de los valores óptimos de las funciones objetivos proporcionados por la rutina AMPL expuesta anteriormente.

Coeficiente C para U_1, U_2 y U_3 y tamaños de muestra de conglomerados $n = 3, 4, 5$. En todos los casos, en cada conglomerado se extraen 10 unidades finales.

n	U_1	U_2	U_3
3	78.21 %	66.68 %	34.32 %
4	87.08 %	75.39 %	38.59 %
5	91.64 %	80.10 %	42.78 %

Podemos observar como en todos los casos se ha obtenido una evidente reducción de la varianza lo que manifiesta que la metodología estudiada en este trabajo se presenta como una alternativa prometedora.

REFERENCIAS

Chambers, R. L. y Dunstan, R. (1986). «Estimating distribution functions from survey data». *Biometrika*, 73, 597-604.

Chambers, R. L., Dorfman, A. H. y Hall, P. (1992). «Properties of stimators of the finite population distribution function». *Biometrika*, 79, 577-582.

Fernández, F. R. y Mayor, J. A. (1995). *Muestreo en poblaciones finitas: curso básico*. Barcelona: E.U.B.

Fourer, R., Gay, D. M. y Kernighan, B. W. (1993). *AMPL. A Modeling Language for Mathematical Programming*. Danvers, Massachusetts: Boyd & Fraser Publishing Company.

Hill, B. M. (1968). «Posterior distribution of percentiles: Bayes' theorem for sampling from a population». *Journal of the American Statistical Association*, 63, 677-691.

Kuk, A. Y. C. (1988). «Estimation of distribution functions and medians under sampling with unequal probabilities». *Biometrika*, 75, 97-103.

Kuk, A. Y. C. y Mak, T. K. (1989). «Median estimation in the presence of auxiliary information». *Journal of the Royal Statistical Society*, Series B, 51, 261-269.

Rao, J. N. K., Kovar, J. G. y Mantel, H. J. (1990). «On estimating distribution functions and quantiles from survey data using auxiliary information». *Biometrika*, 77, 365-375.

Rao, J. N. K. (1994). «Estimating totals and distributions functions using auxiliary information in the estimation stage». *Journal of Official Statistics*, 10, 153-166.

Sedransk, J. y Meyer, J. (1978). «Confidence intervals for the quantiles of a finite population: simple random and stratified simple random sampling». *Journal of the Royal Statistical Society*, Series B, 40, 239-252.

Woodruff, R. S. (1952). «Confidence intervals for medians and other position measures». *Journal of the American Statistical Association*, 47, 635-646.

ENGLISH SUMMARY

ESTIMATING DISTRIBUTIONS FUNCTIONS USING APPROPRIATE SAMPLING DESIGNS IN TWO STAGE CLUSTER SAMPLING

J. A. MAYOR GALLEGO
M. MARTÍNEZ BLANES
Universidad de Sevilla*

In order to estimate the distribution function of a variable defined over a finite population, we can use a sampling strategy defined by the Horvitz-Thompson estimator and a sampling design. The variance of this estimation is a real function whose minimization in a suitable criterion let us find some desirable properties of the appropriate sampling designs.
In this paper we use the $||\cdot||_1$ norm of the variance function as a minimization criteria. This way, under the hypothesis of uniform sampling design, that is to say, with equal first order inclusion probabilities, we study a procedure to obtain appropriate designs under two stage cluster sampling.

Keywords: Sample survey, sampling design, distribution function estimation, cluster sampling

AMS Classification (MSC 2000): 62D05

* Dpto. de Estadística e Investigación Operativa. Universidad de Sevilla. Facultad de Matemáticas. c/ Tarfia s/n, 41012 Sevilla, España. E-mail:mayor@cica.es

– Received March 2001.
– Accepted December 2001.

Usually, the theory of sampling from finite populations is centered on the point estimation of some parameters as the finite population means, variances and ratios. In this paper, we consider the estimation of a functional parameter, the distribution function in relation to a numerical variable, defined over the population.

In literature we can find different approaches to this estimation problem. Chambers and Dunstan (1986) assume a model-based approach to develop an estimating procedure. Kuk (1988) studies several estimators of the distribution function under sampling with unequal probabilities, proportional to an auxiliary variable and Rao, Kovar and Mantel (1990) by means of the auxiliary information. In the same lines, we also can cite the papers of Chambers *et al.* (1992) and Rao (1994).

We propose an alternative approach based on the application of an average-type criterion to the mean square error of the distribution function estimation, in order to find the more appropriate selection probabilities of the clusters.

Let us consider a finite population $U = \{1, 2 \ldots N\}$ and let Y denote the numerical survey variable of interest. Let Y_i be the value of Y for the ith population element, with ordered values $0 \leq Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(N)}$. The aim is to estimate the distribution function of the Y variable,

$$F(t) = \frac{1}{N} \sum_{i \in U} I_{[Y_i, +\infty)}(t)$$

where $I_{[Y_i, +\infty)}(t),\ i \in U$ are the indicator functions of the $[Y_i, +\infty)$ intervals.

If we assume that s is a sample obtained from U whit a sampling design $(S, p(\cdot))$, and $\widehat{F}(t)$ is an estimator of $F(t)$, the classical way to measure the precision of this estimator is to study of the variance,

$$V[\widehat{F}(t)] = \sum_{s \in S} (\widehat{F}(t) - F(t))^2 \, p(s)$$

Note that the variance is a real function with different values depending on t, therefore it is not possible to use this function for a direct comparison. An alternative way to evaluate the discrepancy between $F(t)$ and $\widehat{F}(t)$ is to apply an average type criterion over the variance, considering the quantity,

$$\|V[\widehat{F}(t)]\|_1 = \int_{Y_{(1)}}^{Y_{(N)}} V[\widehat{F}(t)] \, dt$$

as such, we can search the more appropriate sampling designs minimizing $\|V[\widehat{F}(t)]\|_1$.

In section 2 of this paper we apply this approach for a sampling strategy under two-stage cluster sampling, supposing uniform sub-sampling into the clusters. Thus, if we assume that sub-sampling is performed by means of sampling designs $d_i,\ i \in U_c$, we obtain the following result in relation to the variance of the estimation.

Theorem. *If the sampling designs $d_i,\ i \in U_c$ are uniform, with respectively fixed sizes n_i, we have,*

$$\begin{aligned}\|V[\widehat{F}(t)]\|_1 \quad &= \frac{1}{N^2}\sum_{i,j\in U_c}\sum \frac{\pi_{ij}^c}{\pi_i^c\pi_j^c}\sum_{k\in C_i,l\in C_j}\sum (Y_N - \max\{Y_k,Y_l\}) \\ &\quad -\frac{1}{N^2}\sum_{i,j\in U_c}\sum\sum_{k\in C_i,l\in C_j}\sum (Y_N - \max\{Y_k,Y_l\}) \\ &\quad +\frac{1}{N^2}\sum_{i\in U_c}\frac{1}{2\pi_i^c}\left(\sum_{k,l\in C_i}\sum |Y_k - Y_l| - \frac{N_i^2}{n_i^2}\sum_{k,l\in C_i}\sum \pi_{kl}^i |Y_k - Y_l|\right)\end{aligned}$$

Furthermore, if d_i =SRS$(N_i,n_i),\ i \in U_c$, that is to say, simple random sampling without replacement, the last term becomes,

$$\frac{1}{N^2}\sum_{i\in U_c}\frac{N_i}{\pi_i^c}\left(\frac{N_i}{n_i}-1\right)\frac{1}{2N_i(N_i-1)}\sum_{k,l\in C_i}\sum |Y_k - Y_l|$$

□

In order to reduce the sampling error, we develop in sections 3 and 4 a practical procedure based on linear programming, to compute the second order inclusion probabilities of the more appropriate sampling designs over the cluster population, under the super-population model,

$$\boxed{Y_k = \alpha + \beta X_i + \varepsilon_k, \quad \beta > 0, \quad E_s[\varepsilon_k] = 0 \qquad \forall k \in C_i,\ \forall i \in U_c}$$

where X is an auxiliary variable, entirely controlled.

Finally, we have applied this procedure to three artificial populations with different structures. These populations have $M = 20$ clusters, and every cluster contains $N_i = 100$ elements. The sample sizes in the first stage are $n = 3,4,5$ cluster. In the second stage, $n_i = 10$ elements are drawn by means of simple random sampling.

For the population U_1, the auxiliary variable is $X_i = 20 + i,\ i = 1,\ldots,20$, and the super-population model is,

$$Y_k = 10 + 2X_i + \varepsilon_k \qquad \forall k \in C_i,\ \forall i \in U_c$$

with $\varepsilon_k \sim N(0, 3^2)$, that is to say, normal distribution with expectation $\mu = 0$ and variance $\sigma^2 = 3^2$.

For the population U_2, the auxiliary variable is $X_i = i^2,\ i = 1, \ldots, 20$, and the model,

$$Y_k = 20 + 2X_i + \varepsilon_k \qquad \forall k \in C_i,\ \forall i \in U_c$$

with $\varepsilon_k \sim N(0, 5^2)$.

For the population U_3, the auxiliary variable is $X_i,\ i = 1, \ldots, 20$, with values,

$$\{1, 5, 50, 51, 53, 55, 57, 100, 120, 150, 155, 160, 165, 170, 175, 180, 185, 190, 500, 1000\}$$

with the model,

$$Y_k = 20 + 2X_i + \varepsilon_k \qquad \forall k \in C_i,\ \forall i \in U_c$$

with $\varepsilon_k \sim N(0, 10^2)$.

We compare our methodology (MET) with the simple random sampling (SRS) of clusters, using the following relative efficiency measure,

$$C = 100 \times \frac{\|V[\widehat{F}(t)]\|_1^{\mathrm{SRS}} - \|V[\widehat{F}(t)]\|_1^{\mathrm{MET}}}{\|V[\widehat{F}(t)]\|_1^{\mathrm{SRS}}} \%$$

where $\|V[\widehat{F}(t)]\|_1^{\mathrm{SRS}}$ and $\|V[\widehat{F}(t)]\|_1^{\mathrm{MET}}$ are respectively the $\|\cdot\|_1$ of the variance of the compared methods. This way, we have obtained the following values of C,

n	U_1	U_2	U_3
3	78.21 %	66.68 %	34.32 %
4	87.08 %	75.39 %	38.59 %
5	91.64 %	80.10 %	42.78 %

These results show that we can consider our approach as a promising alternative, in order to reduce the sampling error estimating the finite population distribution function.